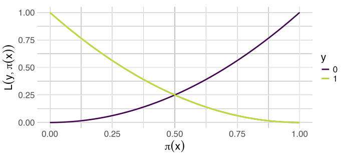
Introduction to Machine Learning

Advanced Risk Minimization
L2/L1 Loss on Probabilities



Learning goals

- Brier score / $L2$ loss on probabilities
- Derivation of risk minimizer
- Optimal constant model
- $L1$ loss on probabilities
- Calibration

BRIER SCORE

- Binary Brier score defined on probabilities $\pi(\mathbf{x}) \in [0, 1]$ and labels $y \in \{0, 1\}$ is $L2$ loss on probabilities

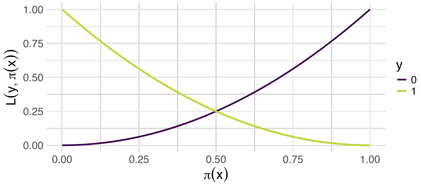
$$L(y, \pi(\mathbf{x})) = (\pi(\mathbf{x}) - y)^2$$

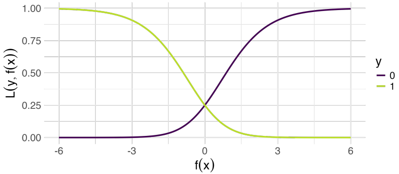
- Despite convex in $\pi(\mathbf{x})$

$$L(y, f(\mathbf{x})) = ((1 + \exp(-f(\mathbf{x})))^{-1} - y)^2$$

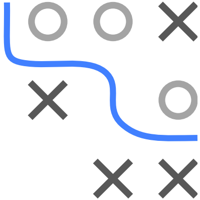
  as composite function not convex in $f(\mathbf{x})$

- Exception would be so-called linear prob. model with $\pi(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$, but that is quite uncommon in ML





BRIER SCORE: RISK MINIMIZER



- Risk minimizer for (binary) Brier score is

$$\pi^*(\tilde{\mathbf{x}}) = \eta(\tilde{\mathbf{x}}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \tilde{\mathbf{x}})$$

- Attains minimum if prediction equals “true” prob $\eta(\mathbf{x})$ of outcome

- Risk minimizer for multiclass Brier score is

$$\pi_k^*(\tilde{\mathbf{x}}) = \eta_k(\tilde{\mathbf{x}}) = \mathbb{P}(y = k \mid \mathbf{x} = \tilde{\mathbf{x}})$$

BRIER SCORE: RISK MINIMIZER

Proof: We only prove the binary case. We need to minimize

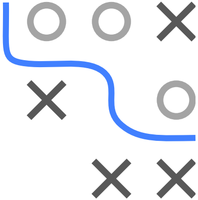
$$\mathbb{E}_x \left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right]$$

which we do pointwise for every $\mathbf{x}$. We plug in the Brier score

$$\begin{aligned}
& \arg\min_c \quad L(1, c)\eta(\mathbf{x}) + L(0, c)(1 - \eta(\mathbf{x})) \\
= & \arg\min_c \quad (c-1)^2\eta(\mathbf{x}) + c^2(1 - \eta(\mathbf{x})) \quad | + \eta(\mathbf{x})^2 - \eta(\mathbf{x})^2 \\
= & \arg\min_c \quad (c^2 - 2c\eta(\mathbf{x}) + \eta(\mathbf{x})^2) - \eta(\mathbf{x})^2 + \eta(\mathbf{x}) \\
= & \arg\min_c \quad (c - \eta(\mathbf{x}))^2
\end{aligned}$$

The expression is minimized for $c = \eta(\mathbf{x})$

BRIER SCORE: OPTIMAL CONSTANT MODEL



- Optimal constant probability model for labels $\mathcal{Y} = \{0, 1\}$ is

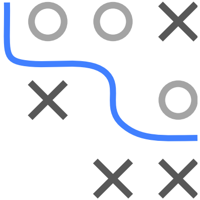
$$\hat{\theta} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\theta) = \arg\min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta \right)^2 = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

- Fraction of class-1 observations in the observed data (directly follows from $L2$ proof for regression)

- Similarly, optimal constant for the multiclass Brier score is

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[y^{(i)} = k]$$

CALIBRATION AND BRIER SCORE



- As Brier score is proper scoring rule, it can be used for calibration
- Prediction $\pi(\mathbf{x}) \in [0, 1]$ called **calibrated** if

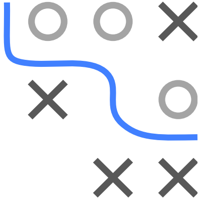
$$\mathbb{P}\big(y = 1 \mid \pi(\mathbf{x}) = p\big) = p \quad \forall p \in [0, 1]$$

- Means: if we predict p, then in $p \cdot 100\%$ of cases we observe $y = 1$ (neither over- nor underconfident)
- Recall RM for Brier score $\pi^*(\mathbf{x}) = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})$. As $\pi^*(\mathbf{x}) = \eta(\mathbf{x})$, optimal predictor satisfies

$$\mathbb{P}\big(y = 1 \mid \pi^*(\mathbf{x}) = p\big) = p$$

i.e., is perfectly calibrated

L1 LOSS ON PROBABILITIES



- Binary L1 loss on probabilities $\pi(\mathbf{x}) \in [0, 1]$ and labels $y \in \{0, 1\}$:

$$L\left(y, \pi(\mathbf{x})\right) = |\pi(\mathbf{x}) - y|$$

- As L1 loss not a proper scoring rule (see part on this), should not necessarily expect good calibration
- Despite convex in $\pi(\mathbf{x})$

$$L\left(y, f(\mathbf{x})\right) = \left|\left(1 + \exp\left(-f(\mathbf{x})\right)\right)^{-1} - y\right|$$

as composite function not convex in $f(\mathbf{x})$